|  |
| --- |
| Harbin Institute of Technology, Shenzhen |
| **Machine Learning Report** |
| House Prices Prediction |
| **Name：张新宇 XinYu Zhang**  **StuId: 19S051009**  **Date: 2019年12月22日** |

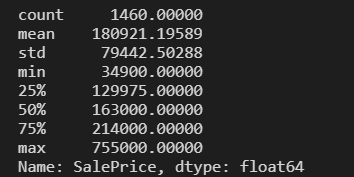
## Chapter 1 Data preprocessing

A machine learning project have many things to do, but its goal is clear and easy to specify. For this project (house price prediction), we focused on regression (or prediction) problem and we want to get a good prediction result. And this derived several problems like what is a ‘good’ prediction, how to deal with high dimensional data, how to fetch useful information from data, and how to combine the information to get a good prediction, this is all very important. And we start from the first question.

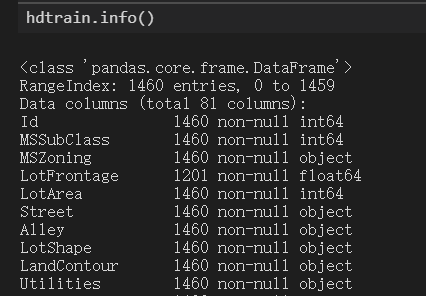
### 1.1 Evaluation index and preface of data

First step: figure out the indices. From the kaggle website we noticed that they use root mean squared error (RMSE) to evaluate the precision of prediction.

From then we start to deal with the data, we use jupyter to load the data and start analysis the relationship between attribute ‘SalePrice’ (The prediction we want get) and other attribute in the training data set. But first we should build a basic impression of statistical significance of the ‘SalePrice’ attribute (Use pandas->describe () method)

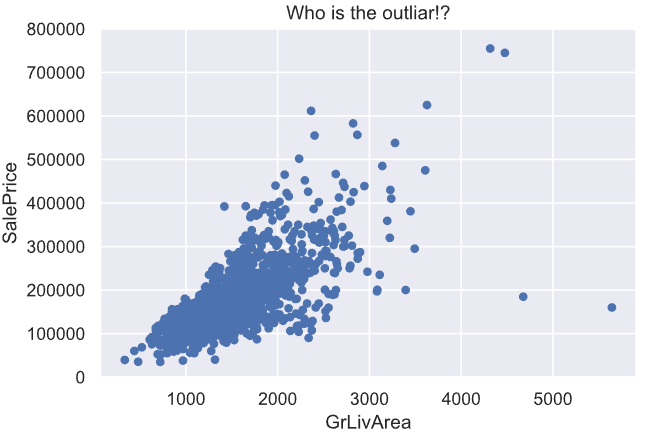


It is a big difference between max and min, but the mean value and 25%, 50%, 75% is relative. So, we may have an assumption about the price’s distribution, but it’s not enough. Before we start to analysis the relationships we should figure out the attribute’s information:

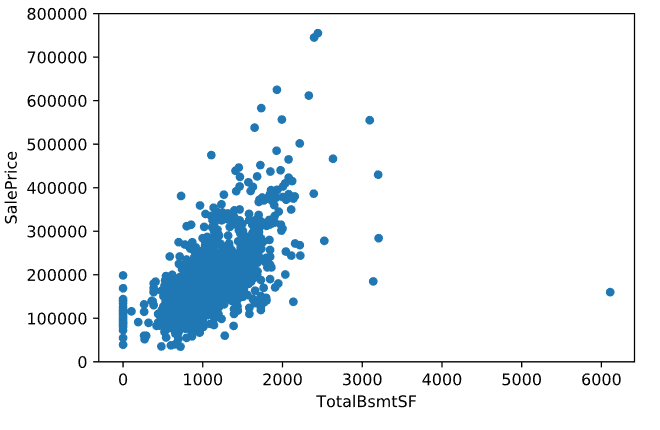


### 1.2 Understand data from plots

From above section the data’s preface is showed, and then we can start to plot some graph to reach a better understanding about relationships between attribute ‘SalePrice’ and other attributes, first we choose ‘GrLivArea’ to start:

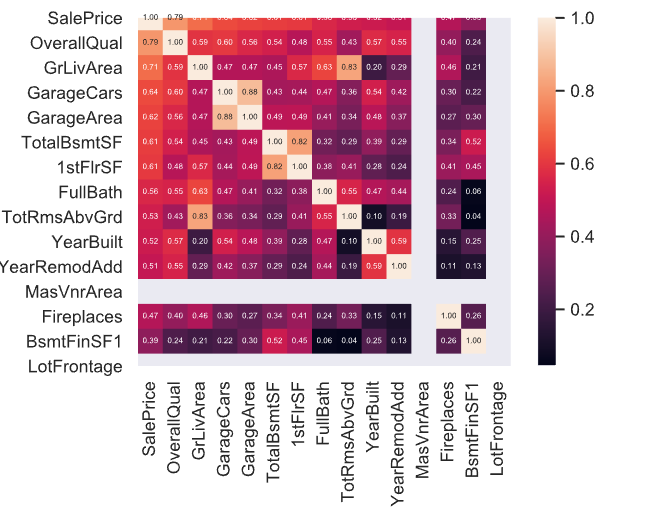


From picture above, we can assume that the relationship between two attributes follows the linear distribution, but we see two points in the right that have a big distance between the main points, although it may be reasonable, it still has some negative influence on our linear assumption. Since then, we decided to remove it from the training set.

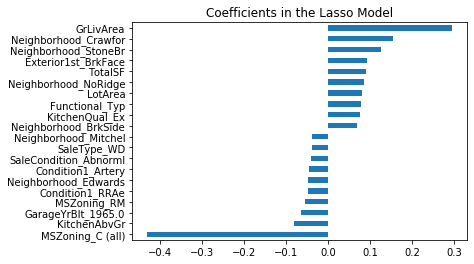


From this picture we can also eliminate some outlier by analyzing and some experience, like the right point bigger than 6000. We remove it from the training set

From above we analysis relationships about two features with ‘SalePrice’, but the number of features is huge that trying to enumerate all the features to analysis is something hard to do. We can use the hot map to observe which attribute is important than any other attributes. We choose 15 important attributes from the training data set:

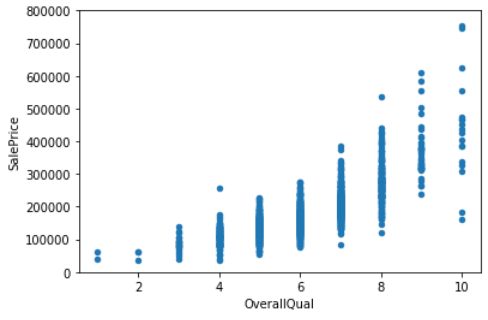


From the hot map we can see a lot of features that have high impact factors on the map, but it still a little bit counter intuition. We try to use lasso to check out which attribute is important and which is not (Noticed that this photo only shows 5 most important and 5 unimportant facts):



Because the ‘OverallQual’, ‘GrLivArea’, ‘GarageCars’, ‘GarageArea’ (Garage\* can be processed into one attribute to make evaluate more easily), TotalBsmtSF (and other SFs), ‘YearBuilt’, etc., we decided to analysis this attributes more careful to get more intuition of the data, this is really deduced our work because the csv file have 81 attributes waited for analysis!

We check the attribute ‘OverallQual’, the plot is shown below:



In this photo, we doesn’t seen much outliers, we decided to cut a little bit and go to analysis next attribute ‘YearBuilt’. From our knowledge we can easily to infer that the completion time of a house is deep related to the price of a house,